Analyzing Data from Completed Experiments

HEPAP
Alexandria

November 8, 2002

Lawrence E. Price
Argonne National Laboratory

Why do it?

a) Provide additional theses from data

b) Analyze data more thoroughly

c) Uncover physics results that would otherwise be unknown

d) Make data available to a wider (taxpaying) public

Questions

a) What criteria should be used for making data available?

b) To whom should the data be provided?

c) What extra work must the collaboration do to provide the data to others?

c) Can outsiders successfully analyze data from today's complex detectors when one needs to understand calibrations, acceptances, trigger biases, dead channels, etc?

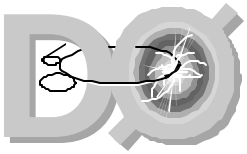
d) Should training be provided as a condition of access?

e) Should the results be validated in any way, such as by a second analysis, godfathers, or other means?

f) How convenient should access to data be made, though GUIs or other means?

g) Who maintains the data and access to it over a period of time?

h) Would systematic errors really be understood by outsiders?



Case Study: Quaero

Making HEP Data Publicly Available

http://quaero.fnal.gov/

Motivation
Data
Algorithm
Examples

Bruce Knuteson
University of Chicago / CDF
UC Berkeley / LBNL / DØ

Quaero

Latin: ***I search for, I seek.***

Designed to address the following problems:

- Easy combination of results from many channels and different experiments
- Rigorous propagation of systematic errors
- Automatic optimization of an analysis
- Reduction of human bias in measurements
- “Full-dimensional” publication of data

→ Many-fold extension of data’s shelf life

→ Saving of LEP data for HEP’s benefit (special case)

Quaero

A General Interface to HEP Data

☐ LEP-II

☐ Aleph ☐ Delphi ☐ L3 ☐ Opal

○ Pythia Input:

○ Signal File:

Browse...

Backgrounds: ☑ qq ☑ e+e- ☑ l+l- ☑ 1ph ☑ 4f ☑ multi-ph ☑ 2ph

☑ TEV-II

☐ DØ ☑ CDF

◉ Pythia Input:

○ Signal File:

Browse...

Backgrounds: ☑ jj ☑ pj ☑ pp ☑ w ☑ z ☑ w ☑ tt

Requestor

Email:

Submit

Quaero is appropriate for high energy collider data

HERA I,II LEP II Tevatron I,II LHC

Each event can be usefully summarized by roughly a dozen numbers

- object type ($e^{\pm}$ $m^{\pm}$ $t^{\pm}$ g $\not{p}$ j b)
- object 4-vectors

Data event:

```
data    1    190.0
e+    45.2   +0.11   0.21
e-    47.3   -0.05   3.56
b     46.0   -0.16   1.71
b     48.2   -0.02   4.90
uncl   0.44   3.3    +0.07   3.97 ;
```

Does it work?

Once Quaero was developed, these 11 analyses were performed as a test of Quaero's correctness and sensitivity.

Results were found to be consistent with (and competitive with) previous results in all cases.

Process	ϵ_{sig}	$\hat{b}$	N_{data}	$\sigma^{95\%} \times \mathcal{B}$
$WW \to e\mu\not{E}_T$	0.14	19.0 ± 4.0	23	1.1 pb
$ZZ \to ee2j$	0.12	19.7 ± 4.1	19	0.8 pb
$t\bar{t} \to e\not{E}_T 4j$	0.13	3.1 ± 0.9	8	0.8 pb
$t\bar{t} \to e\mu\not{E}_T 2j$	0.14	0.6 ± 0.2	2	0.4 pb
$h_{175} \to WW \to e\not{E}_T 2j$	0.02	29.6 ± 6.5	32	11.0 pb
$h_{200} \to WW \to e\not{E}_T 2j$	0.07	66.0 ± 13.8	69	4.4 pb
$h_{225} \to WW \to e\not{E}_T 2j$	0.06	43.1 ± 9.2	44	3.6 pb
$h_{200} \to ZZ \to ee2j$	0.15	17.9 ± 3.7	15	0.6 pb
$h_{225} \to ZZ \to ee2j$	0.15	18.8 ± 3.8	12	0.4 pb
$h_{250} \to ZZ \to ee2j$	0.17	18.1 ± 3.7	18	0.6 pb
$W'_{200} \to WZ \to e\not{E}_T 2j$	0.05	27.7 ± 6.3	29	3.4 pb
$W'_{350} \to WZ \to e\not{E}_T 2j$	0.23	22.7 ± 5.2	27	0.7 pb
$W'_{500} \to WZ \to e\not{E}_T 2j$	0.26	2.1 ± 0.8	2	0.2 pb
$Z'_{350} \to t\bar{t} \to e\not{E}_T 4j$	0.11	18.7 ± 4.0	20	1.1 pb
$Z'_{450} \to t\bar{t} \to e\not{E}_T 4j$	0.14	18.7 ± 4.0	20	0.9 pb
$Z'_{550} \to t\bar{t} \to e\not{E}_T 4j$	0.14	3.8 ± 1.0	2	0.3 pb
$Wh_{115} \to e\not{E}_T 2j$	0.08	37.3 ± 8.2	32	2.0 pb
$Zh_{115} \to ee2j$	0.20	19.5 ± 4.1	25	0.8 pb
$LQ_{225}\overline{LQ}_{225} \to ee2j$	0.33	0.3 ± 0.1	0	0.07 pb

(Table from Quaero Phys. Rev. Lett.)

DØ chose to make data available with general scope and limited internal review

Quaero Policy

- Any "interesting" Quaero result will be reviewed by a DØ Quaero Review Board.
  - A Quaero result is "interesting" if an excess of data over background of more than 2.0 standard deviations is found.
  - If an interesting result is found, the requestor is notified that his request is under review, and the result of the request is sent to the review board.
  - If a fault is found the fault is rectified, the request is re-run, and the new result is sent to the requester (along with an explanation).
  - If the "interestingness" is not deemed to be due to any fault, the result is sent to the requester.
- In all cases the requester is free to publish the Quaero result in his or her own paper, so long as Quaero is referenced. The appropriate citation, including the Quaero request log number and request date, is included in the email with Quaero's result.

Time/effort to develop Quaero?

Quaero idea originally hatched: December 2000

Quaero made public: June 2001

Total man-effort: 1 postdoc x 6 months

Total required resources: 1 three-year-old Linux PC

The additional time/effort to make the data public using Quaero is negligible

(Compared to the 10^4 person-years to build the DØ detector and understand the data, << 0.01%)

Any experiment wishing to use Quaero needs to provide 4 things:

- Data
  - Object 4-vectors
- Backgrounds
  - Object 4-vectors
- Systematic errors
  - Sources of error & effect on 4-vectors
- Detector simulation
  - (fast or full)

Quaero algorithm overview

(you wish to test a hypothesis H)

- H events are run through the detector simulation
- H, SM, data are partitioned into final states
- Variables are chosen automatically
- Binning is chosen automatically
- A binned likelihood is calculated
- Results from different final states are combined
- Results from different experiments are combined
- Systematic errors are integrated numerically

Quaero returns a single number:

$$\mathcal{L}(\mathcal{H}) = \frac{p(\mathcal{D}|\mathcal{H})}{p(\mathcal{D}|\mathrm{SM})},$$

where H is the hypothesis being tested.

From this you can easily compute anything you want (parameters, limits, . . .)

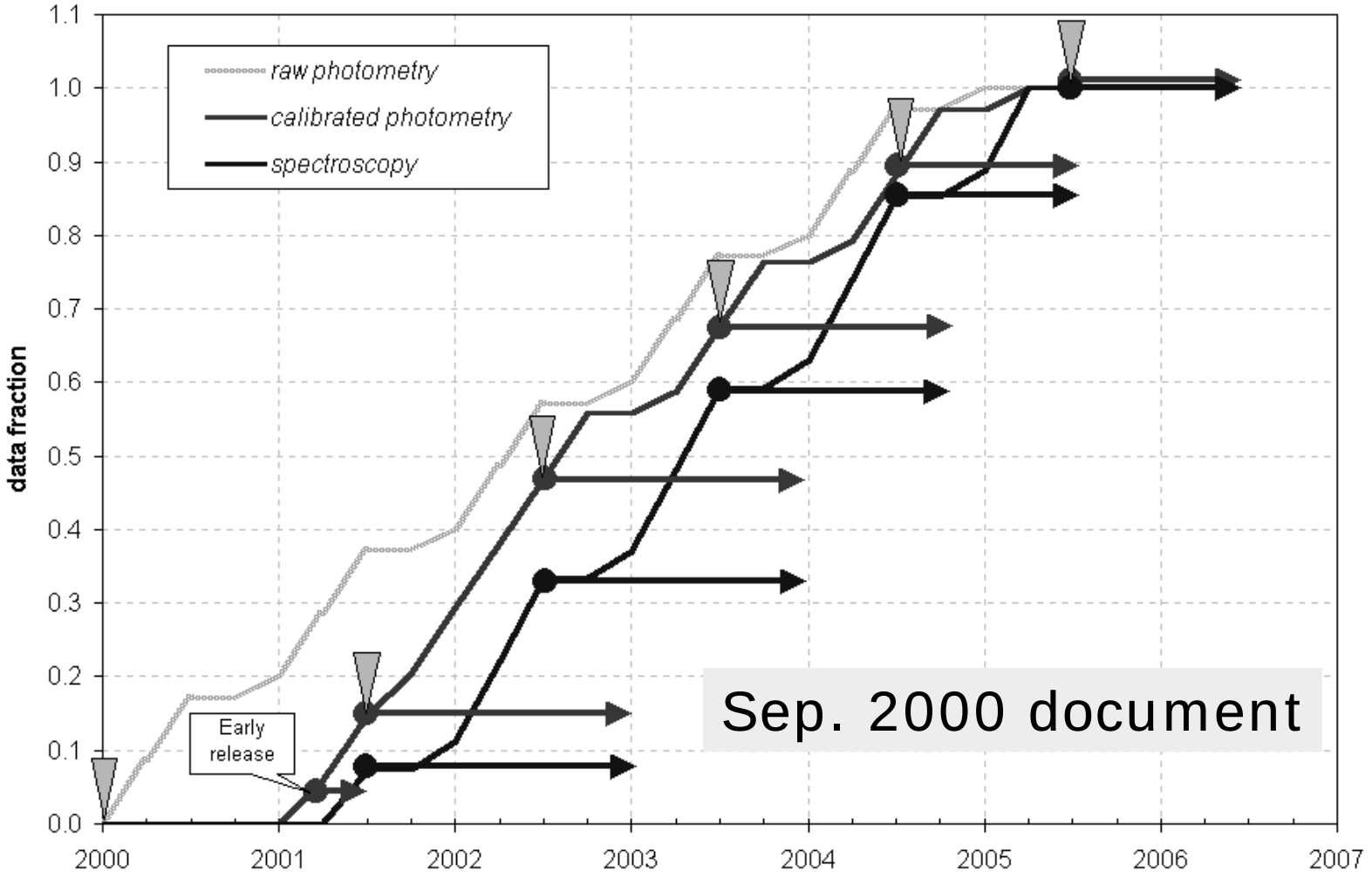
Case Study: SDSS

- **Sloan Digital Sky Survey (2000-2005)**
  - **Survey 25% of sky (10,000 square degrees) in 5 years**
  - **2.5m wide-field telescope in Apache Point, NM**
  - **5 broad color bands (354, 477, 623, 763, 913) nm**
  - **Uniform systematics!**
- **SDSS Science Archive (when complete)**
  - **3 TB of data, 200M objects, stored at Fermilab**
  - **Galaxies → 1M medium res. spectra**
  - **Quasars → 100K medium res. spectra**
  - **Moving object catalog (> 300K objects)**
- **Multiple funding agencies**
  - **Sloan, NSF, NASA, DOE, some foreign**

SDSS Public Data Release

- **NSF Astronomy Division recognized unique importance of SDSS data to astronomy community.**
  - **Charged Astrophysical Research Consortium, which manages SDSS, to develop plan for public data release**
  - **Plan created in 1998, approved in 1999**
- **Release schedule**
  - **“Early release” in mid-2001**
  - **5 yearly releases, first in Jan. 2003 (fig.)**
  - **Release is 18 months after data collection**
  - **Release is 12 months after calibration**
- **Data processing delayed for current data, so public release in 2003 reduces SDSS exclusive access**

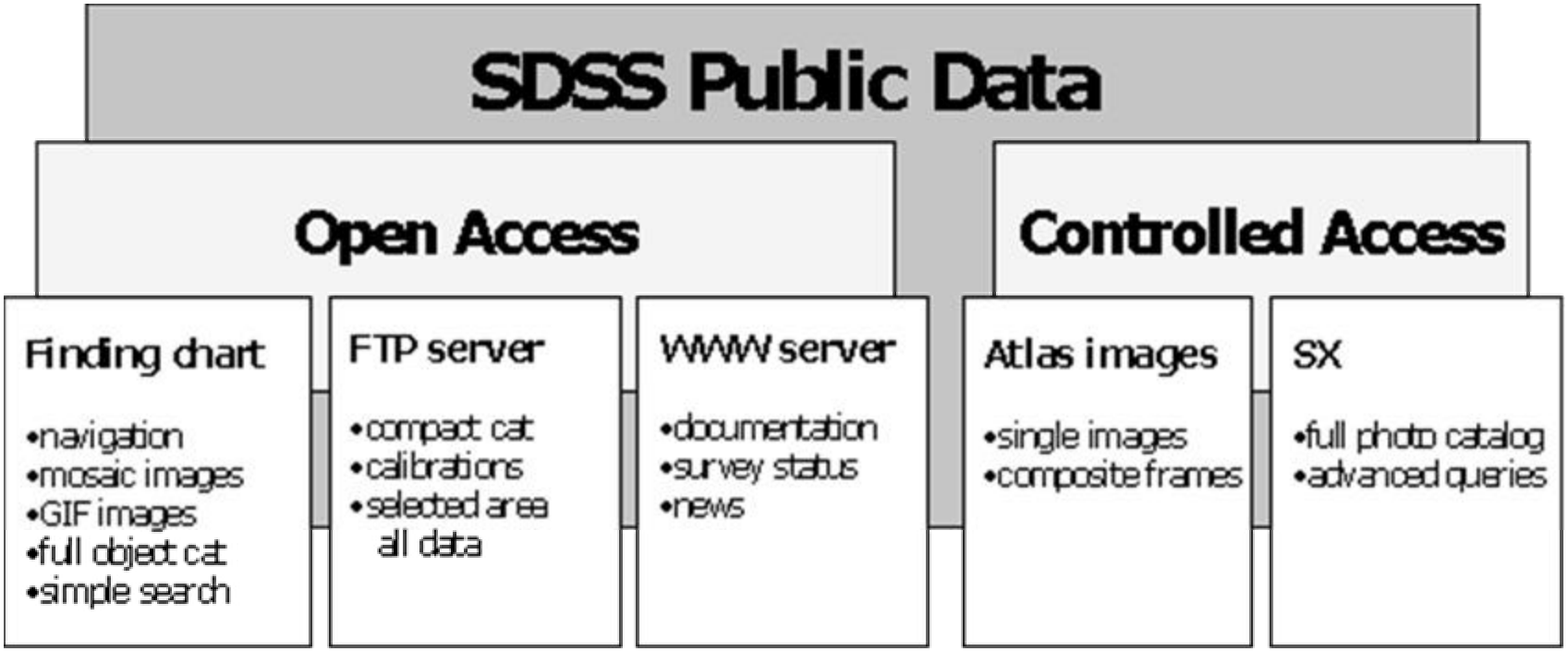
SDSS Data Release Schedule



Details of Access

- **SDSS requested additional funds for access**
  - **Takes real resources**
- **Details for accessing data available on website**
  - **Calibration and data collection procedures/documents**
  - **Descriptions of data, glossaries**
  - **Data reduction programs(?)**
  - **E-mail archives!**
- **Early access data has restrictions due to finite computing resources at Fermilab (fig.)**
  - **Open access: Web-based interface to catalogs**
  - **Controlled access: Full access to Science Archive**
  - **Might merge later if resources deemed adequate**

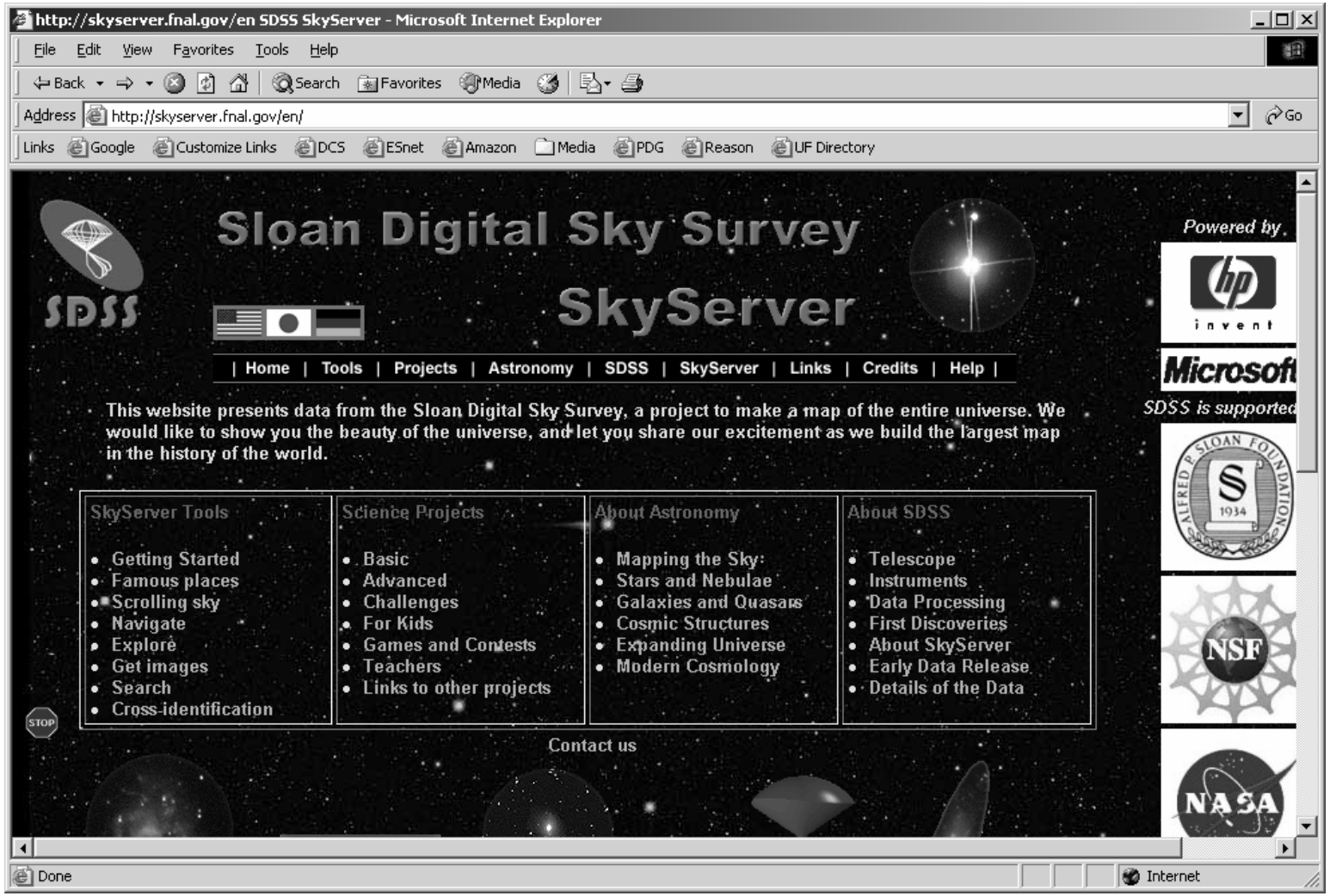
Access to SDSS Data



SDSS SkyServer

- **Provides interactive analyses via web interface**
  - **http://skyserver.fnal.gov/**
  - **Uses fraction of science archive copied to safe area**
  - **80 GB, 14M objects, 50K spectra**
  - **Supported by Microsoft**
- **Provides tools for many science projects**
  - **1 full-time person, telecommutes from Orlando**
  - **High school science teacher helps design projects**
  - **Example: verify Hubble law with large statistics**

SkyServer Web Site



Users of Astronomical Data

Wide public

- Amateur astronomers, high school & college students
- Large number of small queries
- Extra packaging – explanations, glossary, lessons via web

Astronomy community

- Professional astronomers (5K in US, 10K world-wide)
- Medium number of intermediate queries
- Every person using the system at least once a day

Power users

- Small number of very intensive grid computing projects
- Require most of the computing resources
- Use data mirrors at supercomputer centers or TeraGrid

Potential Next Steps….

Collect more data

- **Wait for results from the Quaero implementations at LEP and from further use of D0 data**
- **Learn more about the uses of NASA data by other scientists and the general public**

Request a detailed study

- **HEPAP subcommittee?**

Promote tests and prototypes

- **Encourage Quaero development and a later report**
- **Provide funds for faster Quaero implementation or other data access system (call for proposals?)**

Commission a system for use across HEP

- **Call for proposals?**

Summary

- **NASA example and Quaero system offer successful prototypes**
  - **Comparisons of many D0 analyses by Quaero with earlier analyses by traditional means are very encouraging**
- **Many questions often asked about analysis by "outsiders" are not fully answered yet**
  - **More study could help to clarify some issues**
  - **Real answers will only come by developing and trying a production system on a realistic scale**